Biology-informed Bayesian models for interpretable cancer diagnosis

Stanley E. Lazic

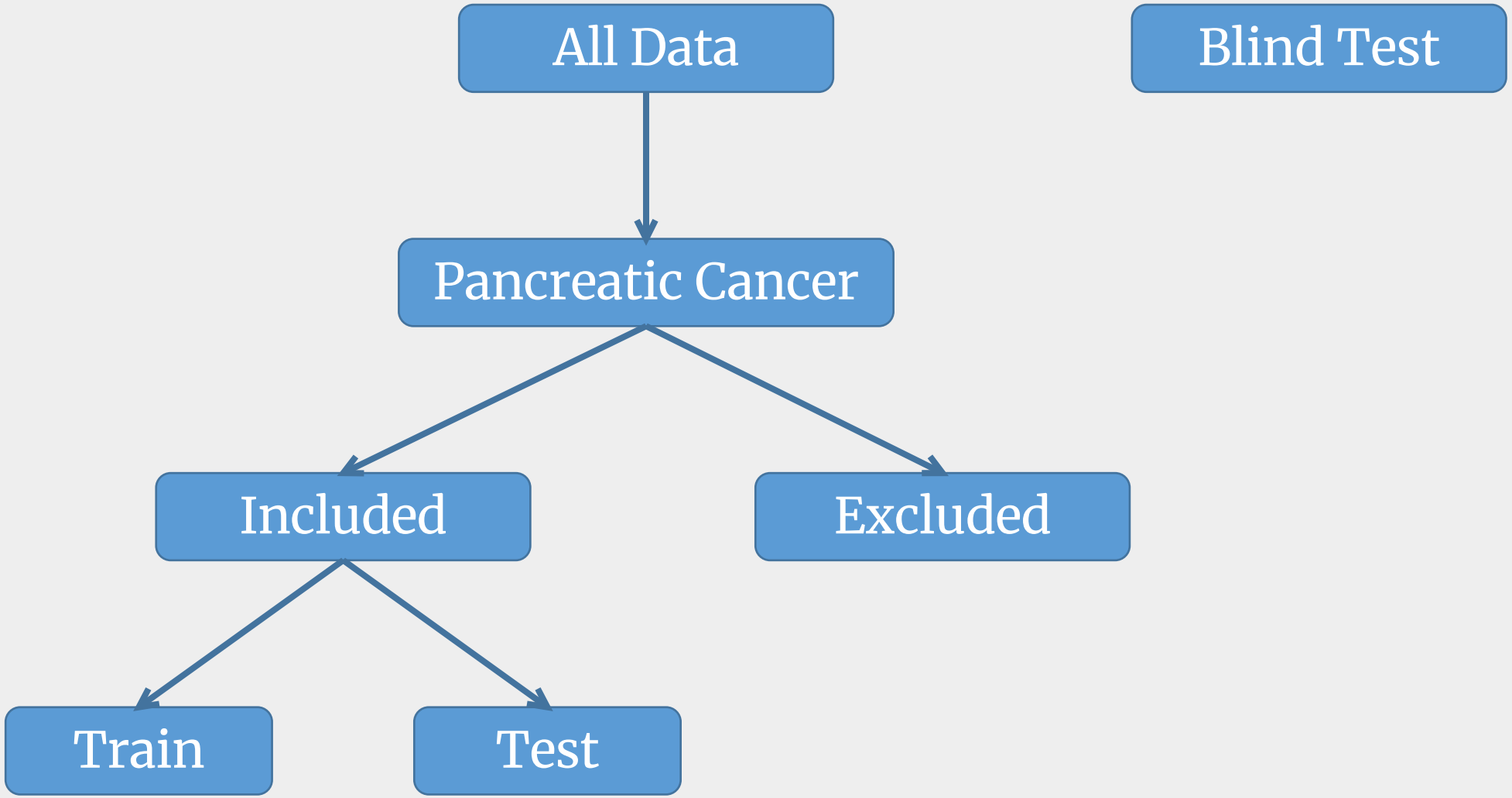
Prioris.ai Inc.

30 May 2025

Aim

Develop a diagnostic prediction model for early-stage pancreatic cancer from protein biomarkers measured in the blood.

The data

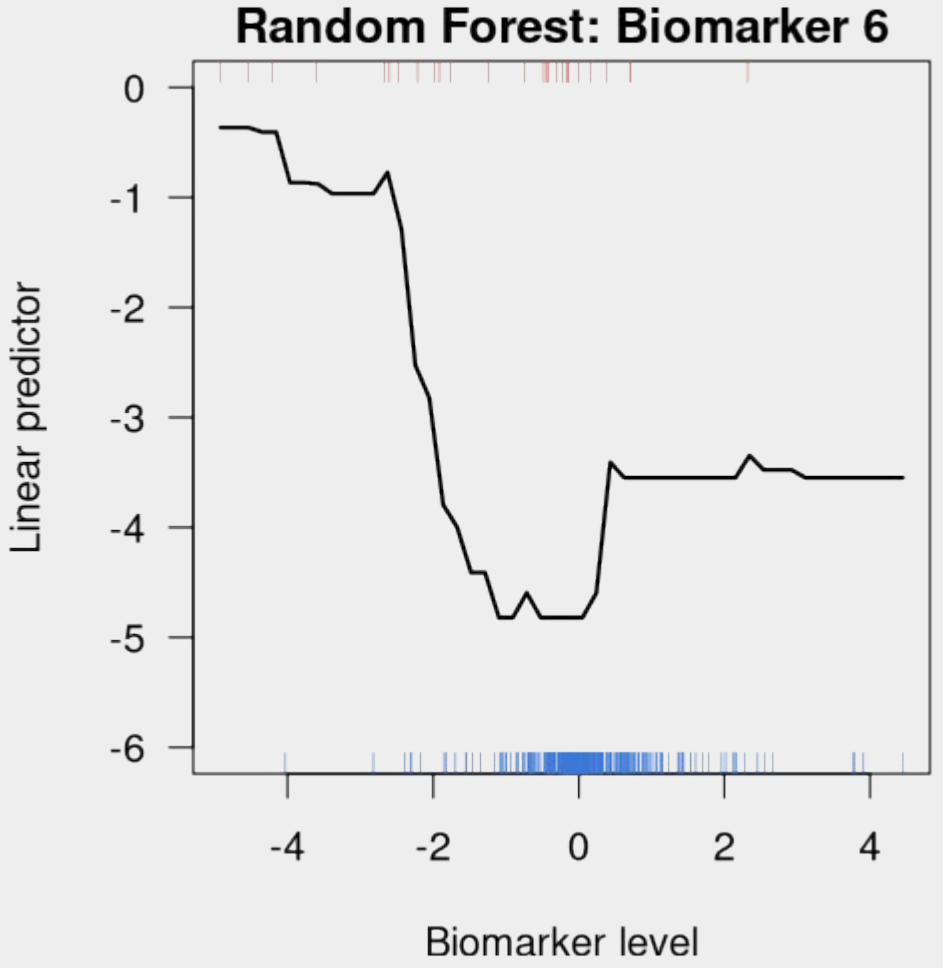

All Data
Blind Test
Pancreatic Cancer
Included
Excluded
Train
Test

Pancreatic cancer data details

- N = 408 (381 Healthy, 27 Cancer)
- Early stage cancer (Stage 1 and 2)
- 186 Males, 222 Females
- Mean age = 62.2 (range: 47 to 76)
- 50 protein biomarkers measured in the blood
- Biomarkers are normalized and standardized
- Case-Control design

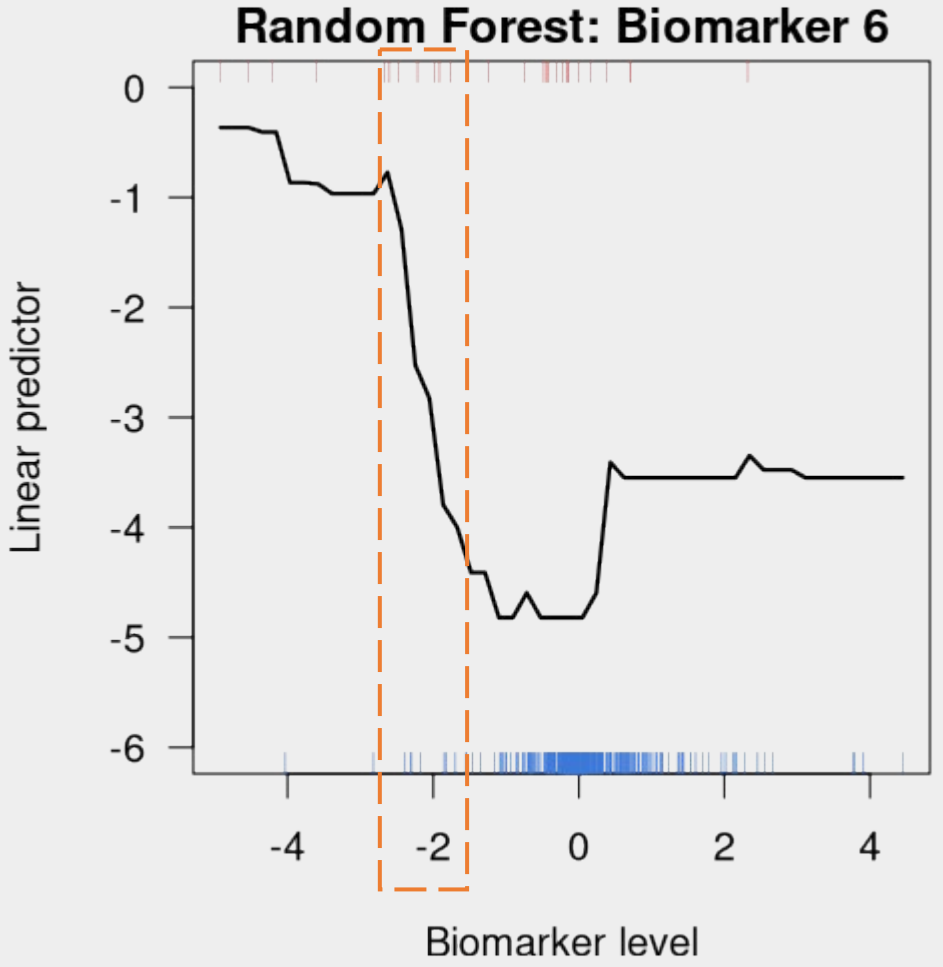
What do we know about the relationship between biomarkers and cancer:

- Direction
- Nonlinearity
- Monotonicity
- Smoothness



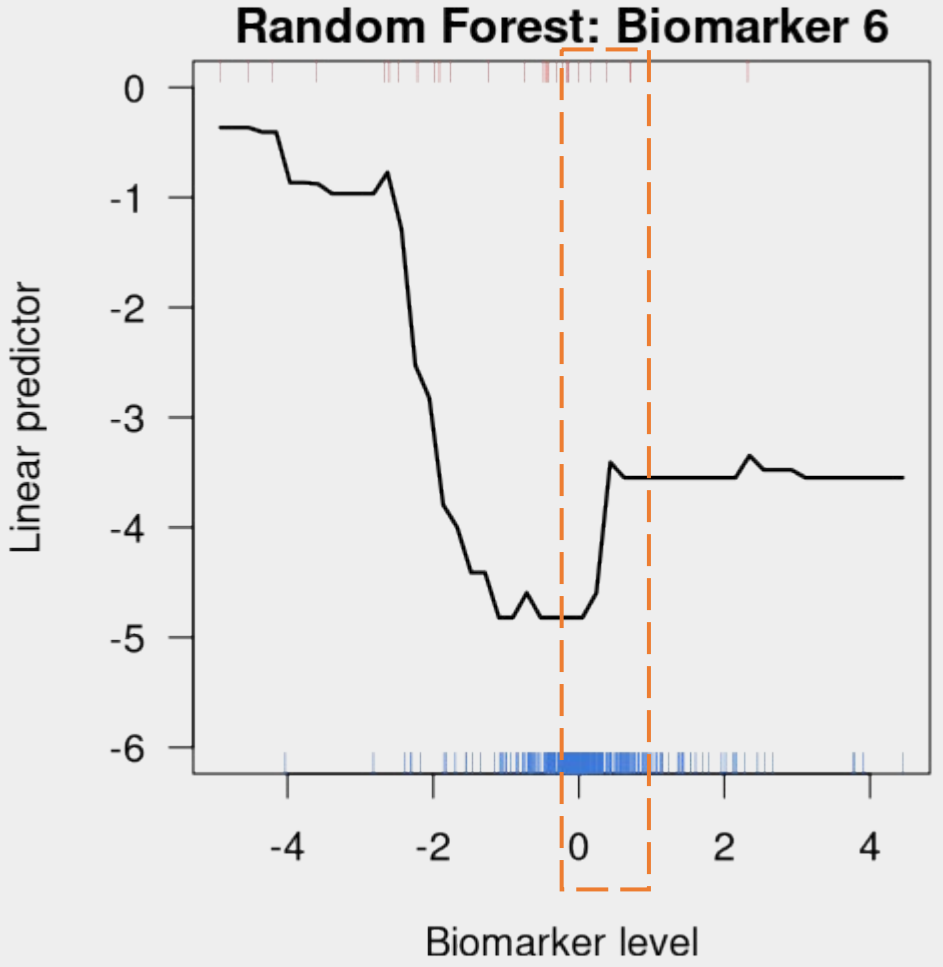
What do we know about the relationship between biomarkers and cancer:

- Direction
- Nonlinearity
- Monotonicity
- Smoothness



What do we know about the relationship between biomarkers and cancer:

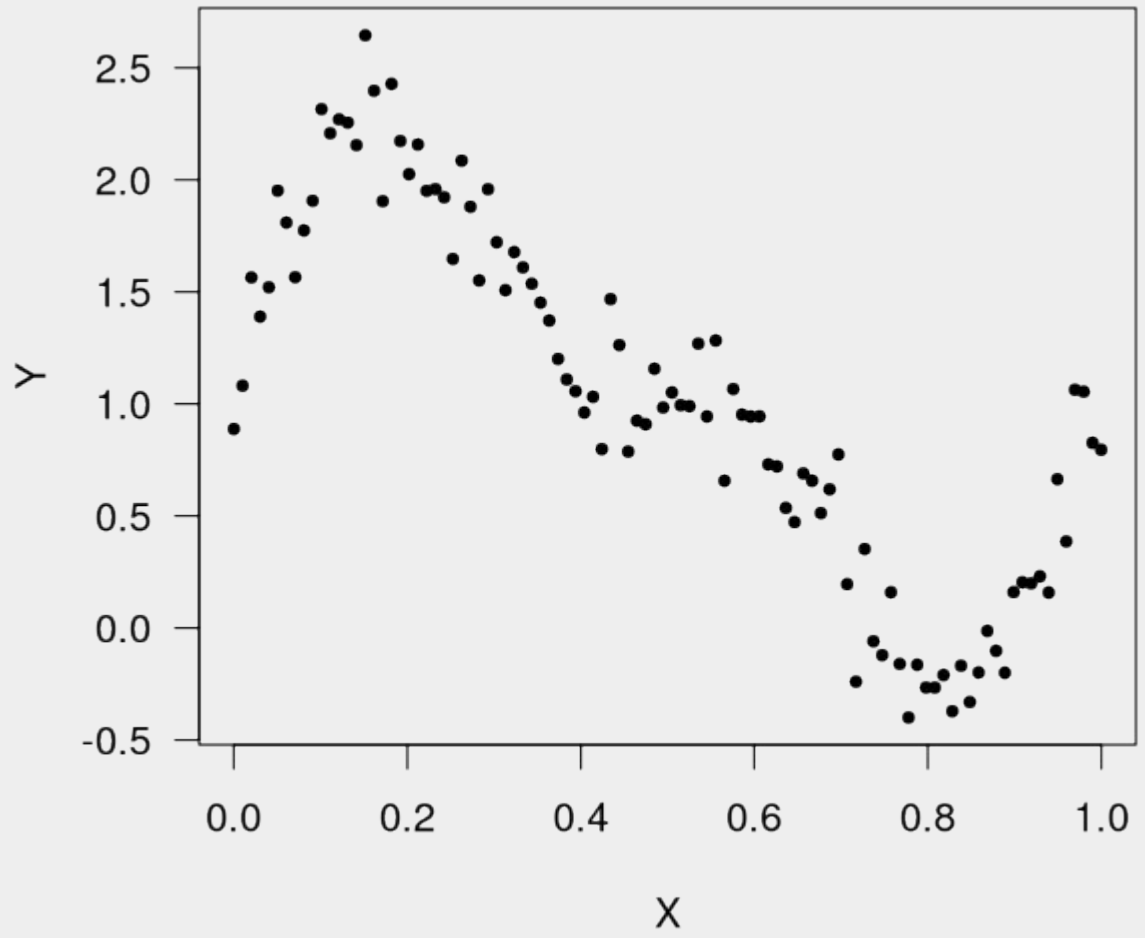
- Direction
- Nonlinearity
- Monotonicity
- Smoothness



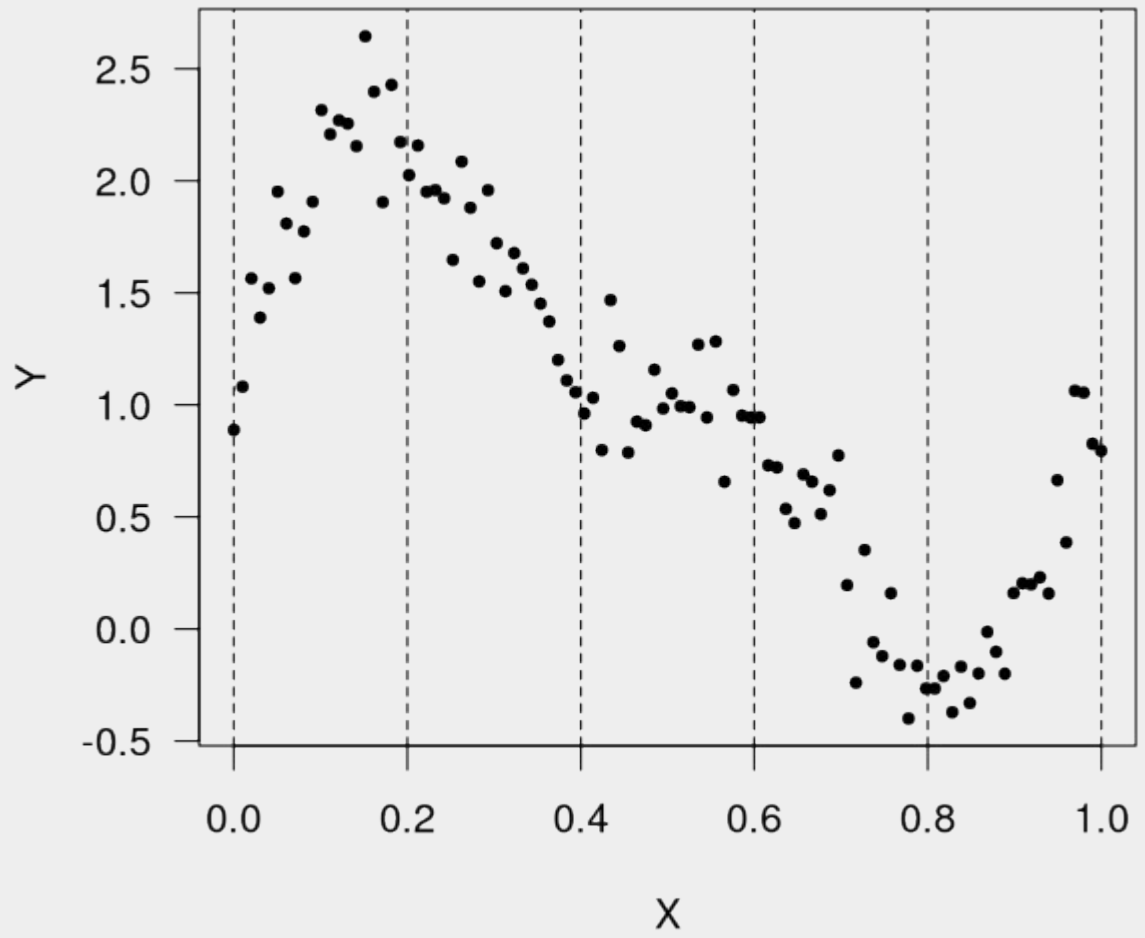
Splines and GAMs

(Generalized Additive Models)

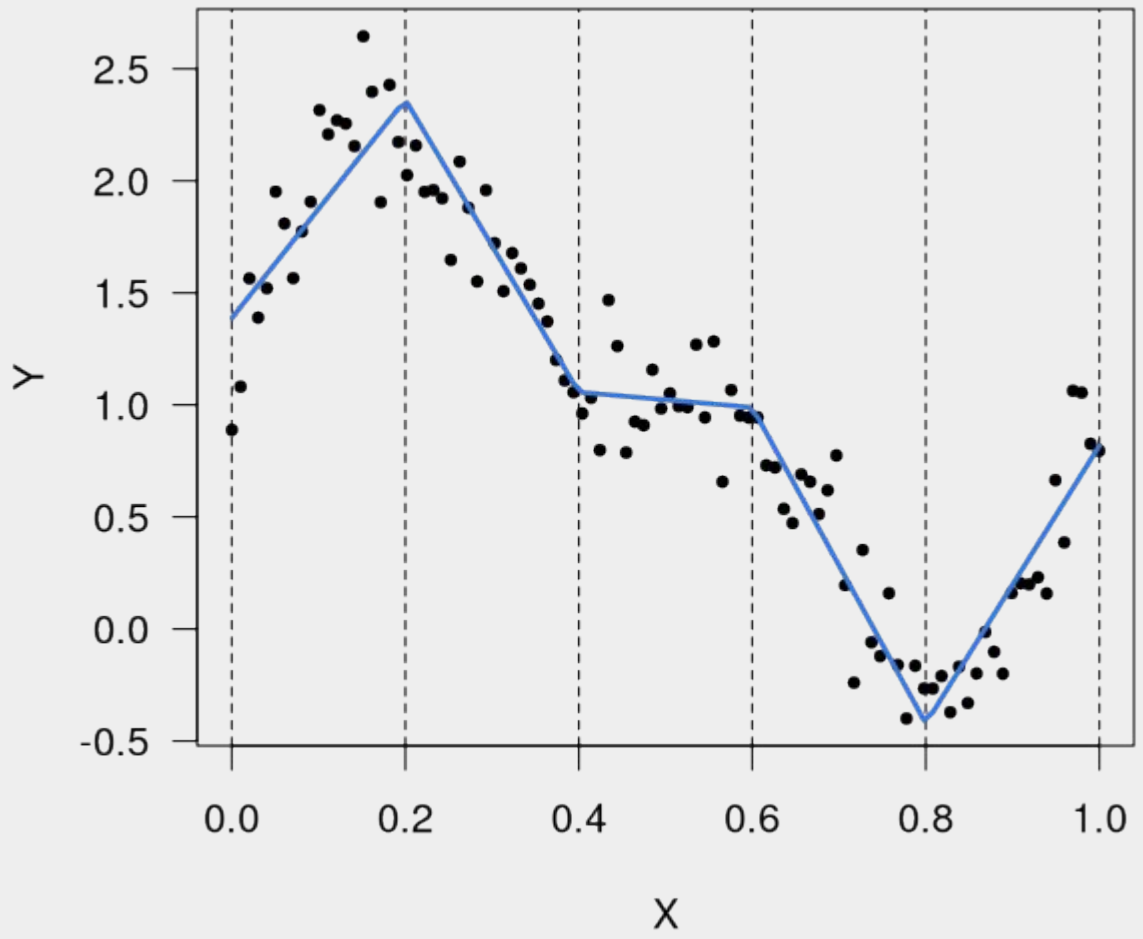
Splines



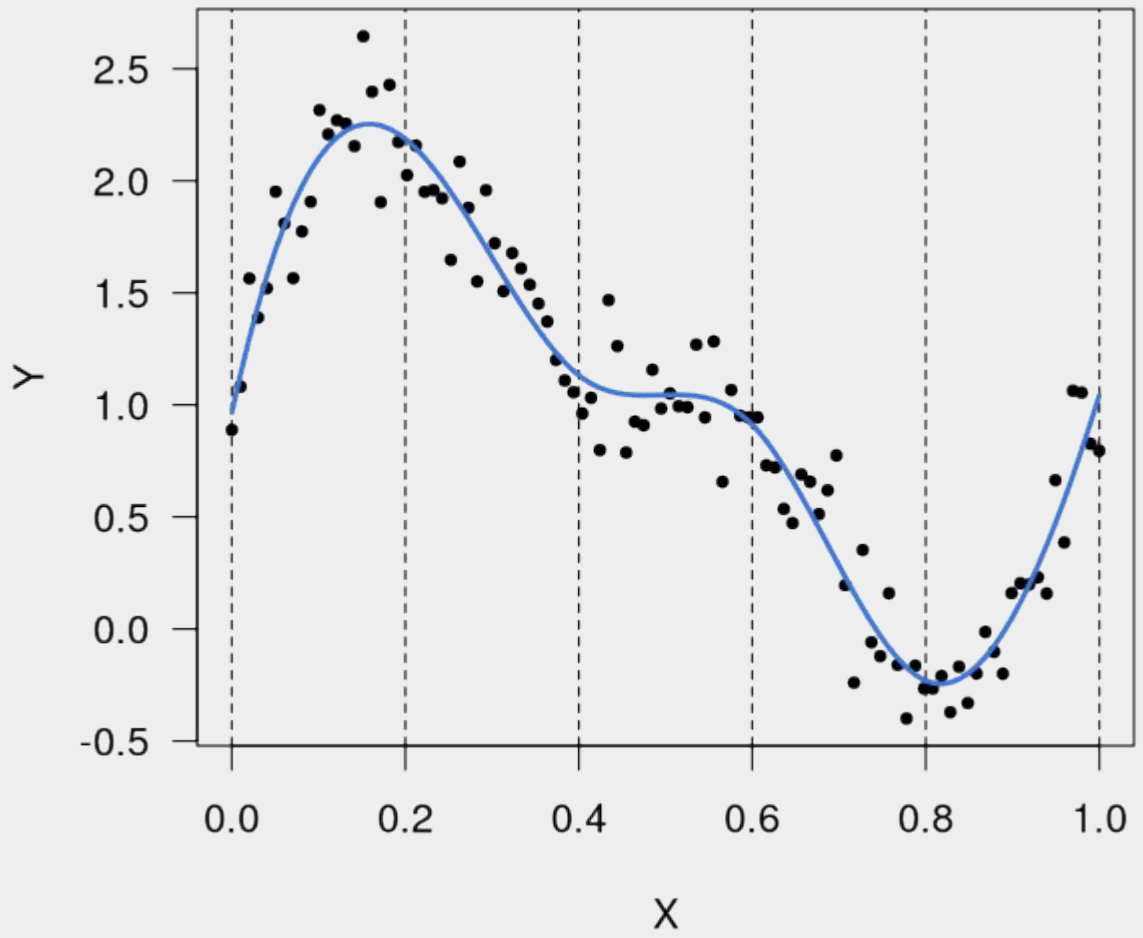
Splines



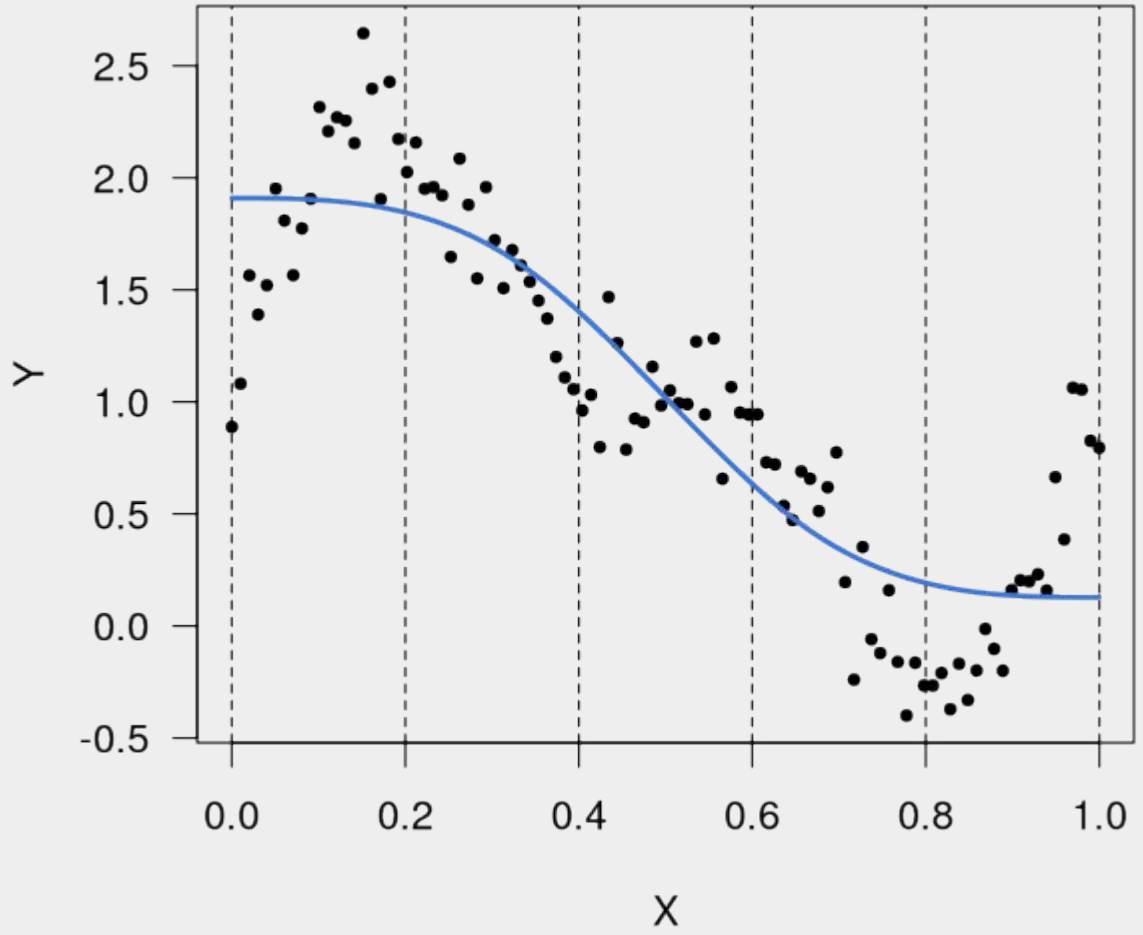
Splines



Splines (smooth and nonlinear)

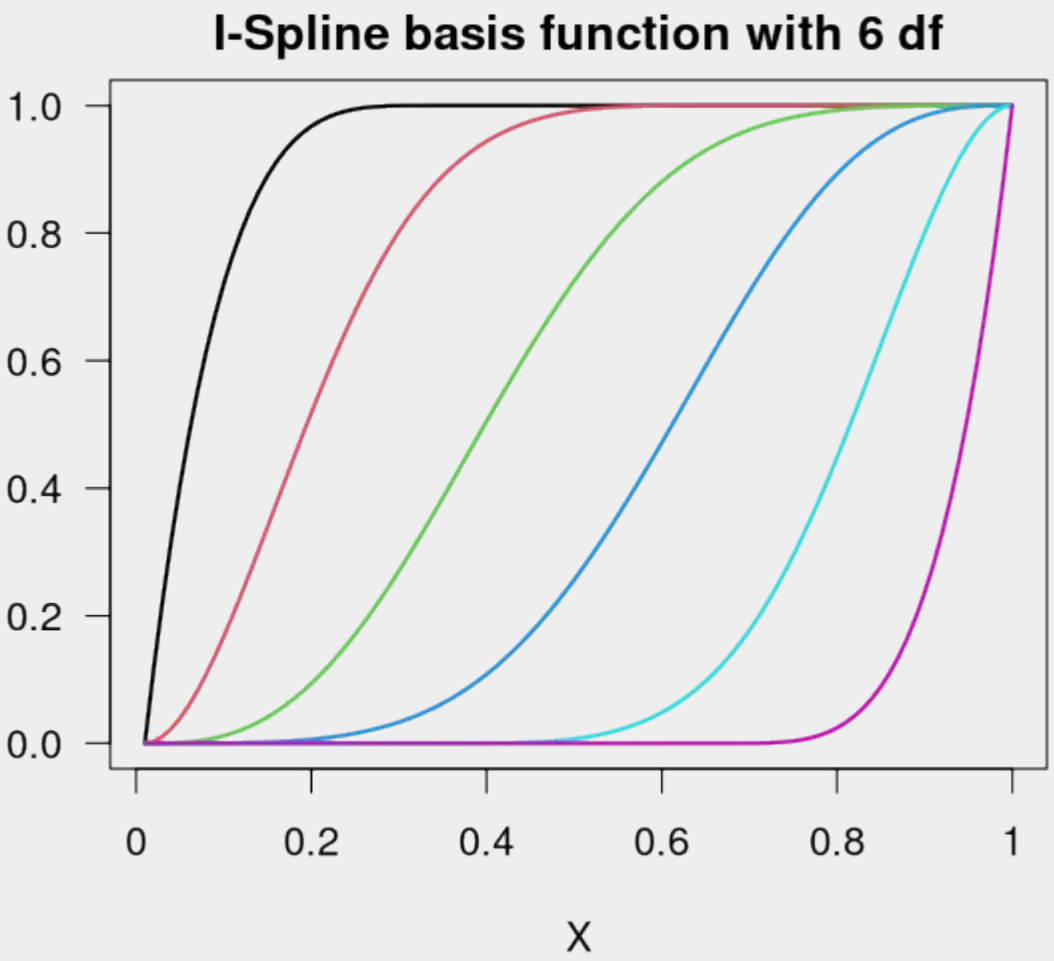


I-Splines (direction and monotonic)



Basis expansion

	1	2	3	4	5	6
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.116	0.003	0.000	0.000	0.000	0.000
3	0.221	0.010	0.000	0.000	0.000	0.000
4	0.317	0.023	0.000	0.000	0.000	0.000
5	0.404	0.039	0.001	0.000	0.000	0.000
6	0.482	0.059	0.002	0.000	0.000	0.000
7	0.552	0.082	0.004	0.000	0.000	0.000
8	0.615	0.108	0.006	0.000	0.000	0.000
9	0.671	0.137	0.008	0.000	0.000	0.000
10	0.720	0.167	0.012	0.000	0.000	0.000
11	0.764	0.199	0.016	0.000	0.000	0.000
12	0.802	0.233	0.021	0.001	0.000	0.000
13	0.836	0.268	0.027	0.001	0.000	0.000
14	0.865	0.303	0.033	0.001	0.000	0.000
15	0.890	0.339	0.041	0.002	0.000	0.000
16	0.911	0.375	0.050	0.002	0.000	0.000



Decisions to be made

- How many knots? → More than you think you need, and regularize.
- Location of knots? → Equally spaced.

Splines and GAMs

One variable:

$$E(y) = \theta_0 + f(x; \theta)$$

Multivariable:

$$E(y) = \theta_0 + f_1(x_1; \theta_1) + \cdots + f_n(x_n; \theta_n)$$

Splines in Stan

```r
X1_train <- splines2::iSpline(biomarker1, df = 6) # N x 6 matrix
```

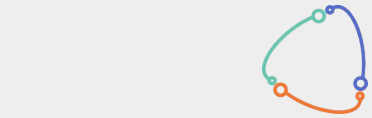
```
data {
  int N;                        // sample size
  int P;                        // num params for each biomarker (bm)

  array[N] int y_train;         // binary outcome
  matrix[N, P] X1_train;        // predictor matrix for bm 1
  matrix[N, P] X2_train;        // predictor matrix for bm 2
}

parameters {
  real b0;                          // intercept
  vector<lower=0>[P] b1;            // param vector for bm 1 (increasing)
  vector<upper=0>[P] b2;            // param vector for bm 2 (decreasing)
  real<lower=0> sigma_penalty1;     // regularization for bm 1
  real<lower=0> sigma_penalty2;     // regularization for bm 2
}
```

Splines in Stan

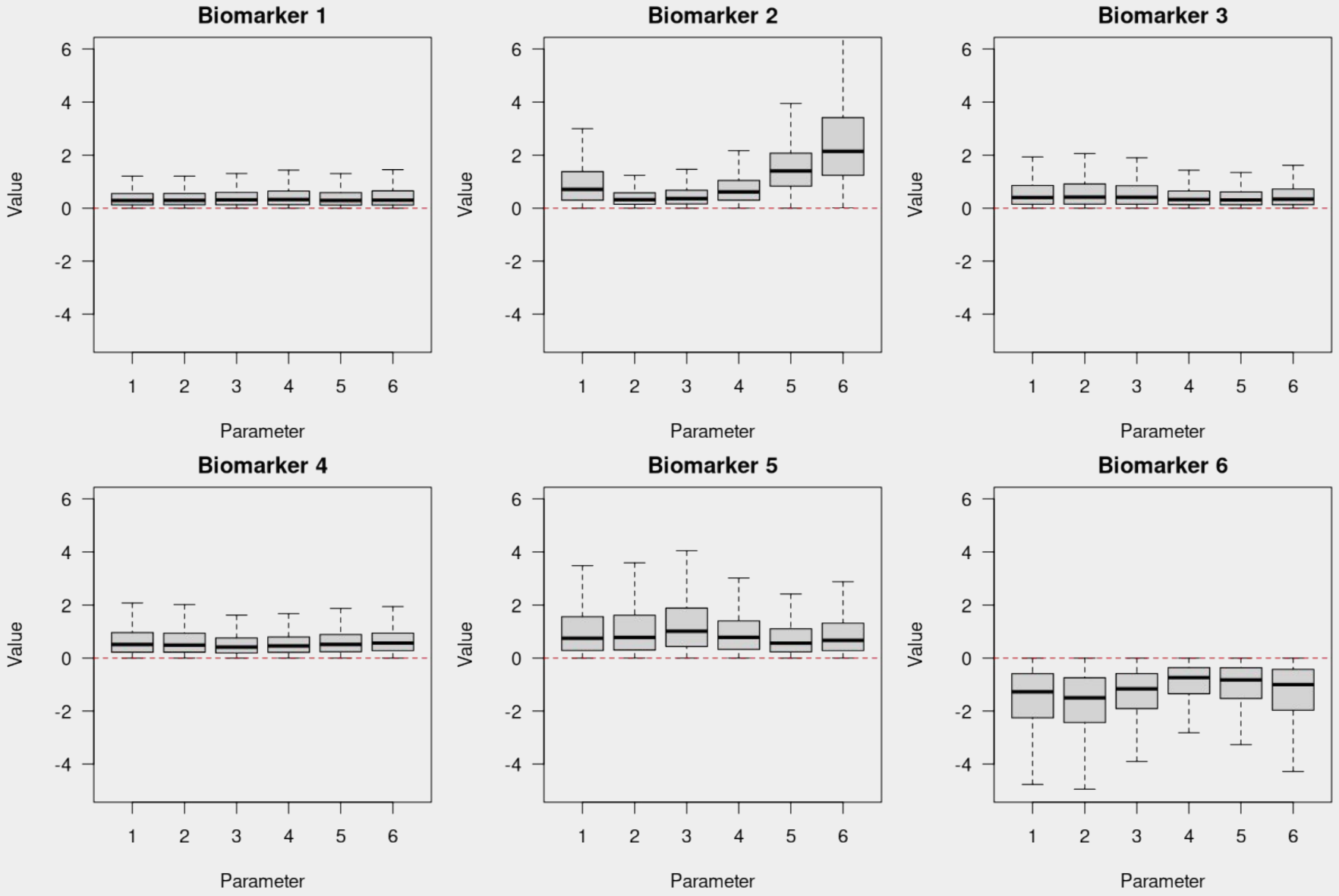
```
transformed parameters {
  // linear predictor
  vector[N] linpred = b0 + X1_train * b1 + X2_train * b2;
}

model{
  // priors
  b0 ~ normal(0, 50);
  b1 ~ normal(0, sigma_penalty1);
  b2 ~ normal(0, sigma_penalty2);
  sigma_penalty1 ~ exponential(0.5);
  sigma_penalty2 ~ exponential(0.5);

  // likelihood
  y_train ~ bernoulli_logit(linpred);
}
```

Splines in Stan

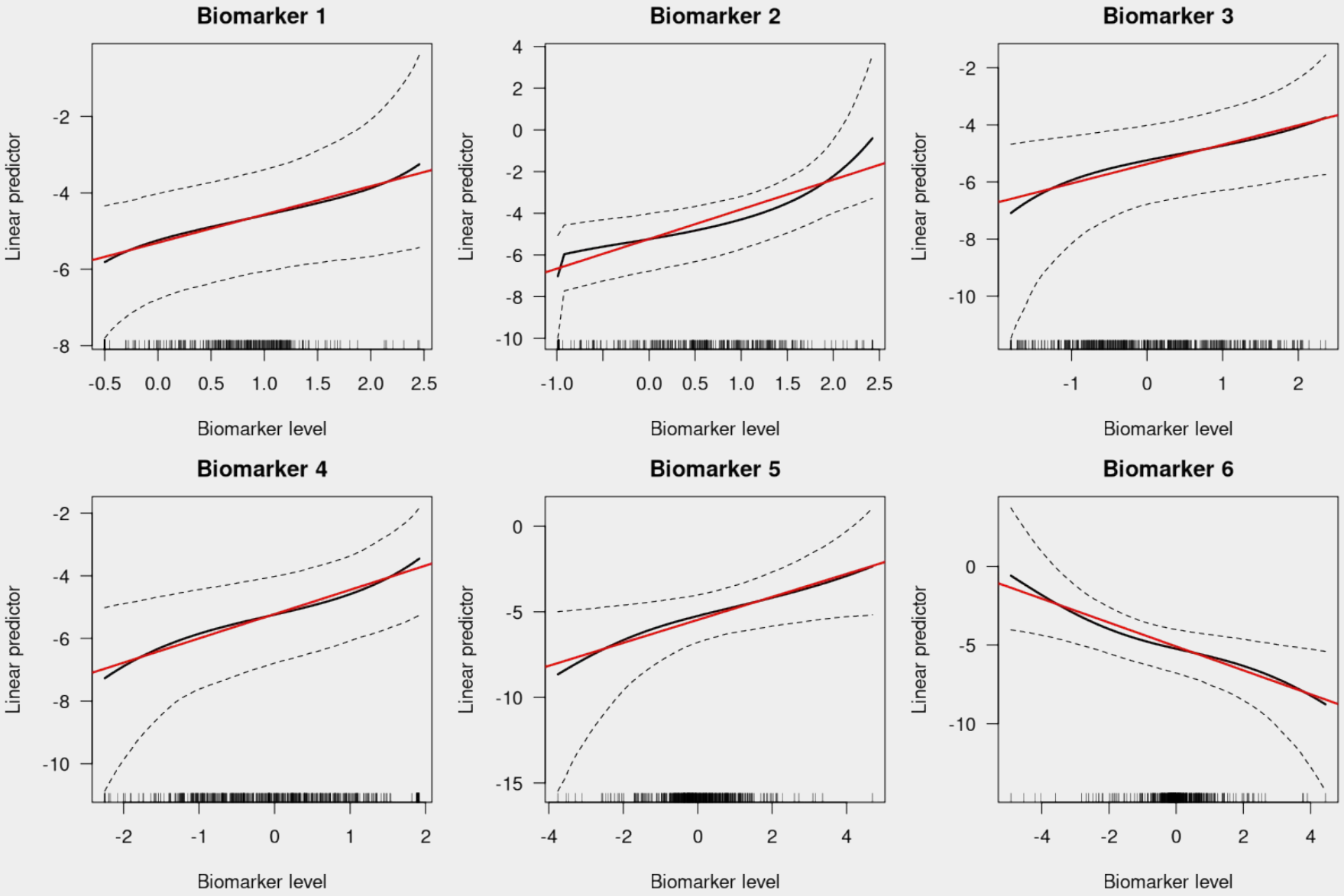
```
generated quantities {
  vector[N] ypred;

   for (i in 1:N){
    // predicted probability of cancer
    ypred[i] = bernoulli_logit_rng(linpred[i]);
  }
}
```



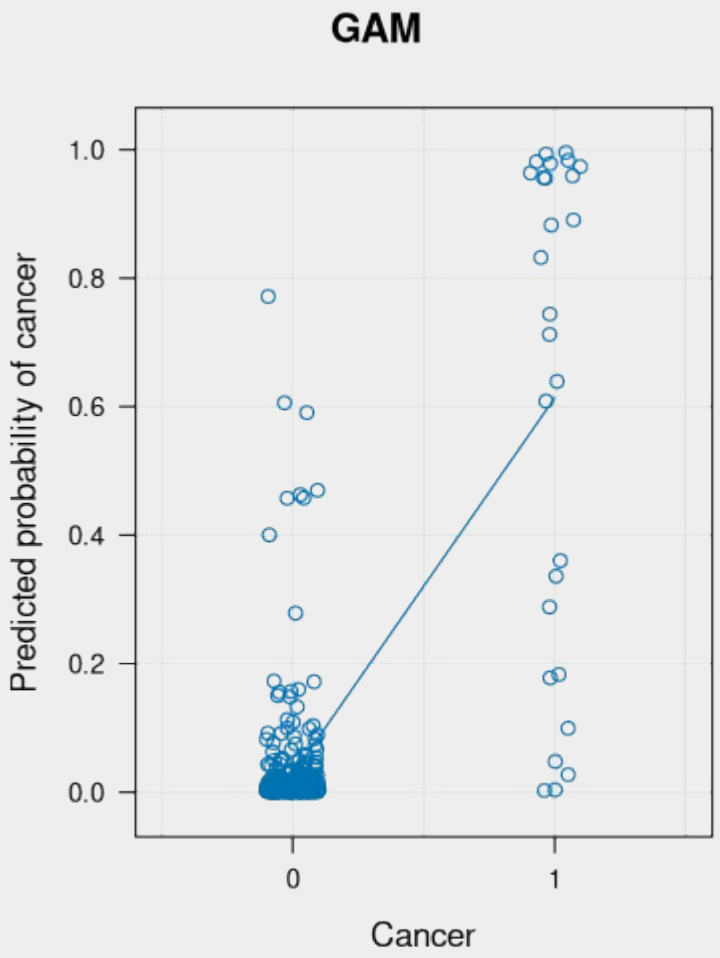
Biomarker parameter estimates

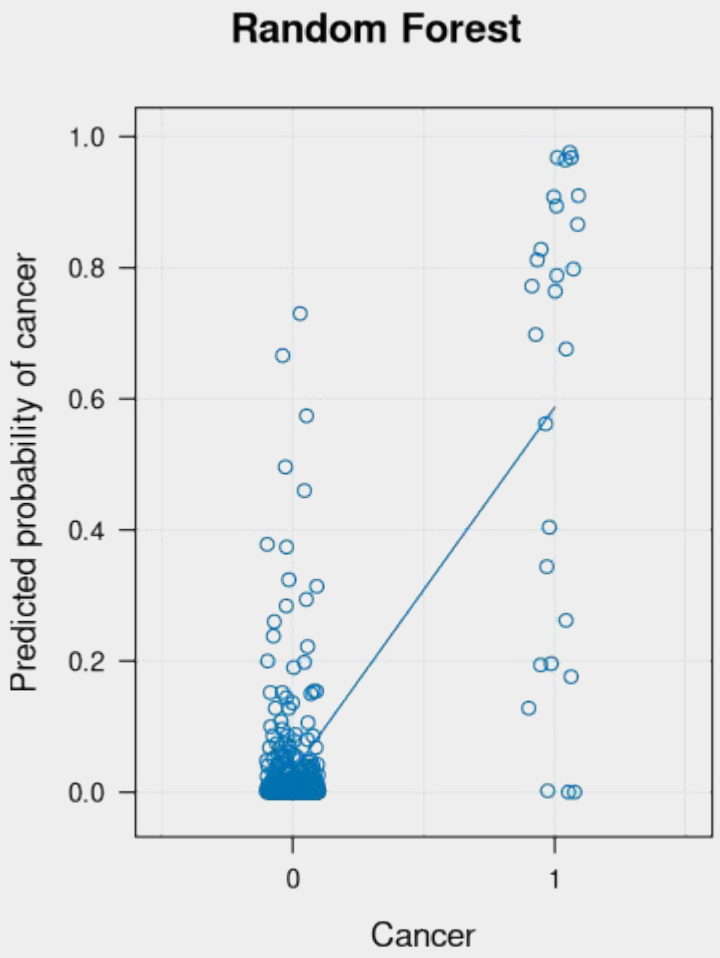


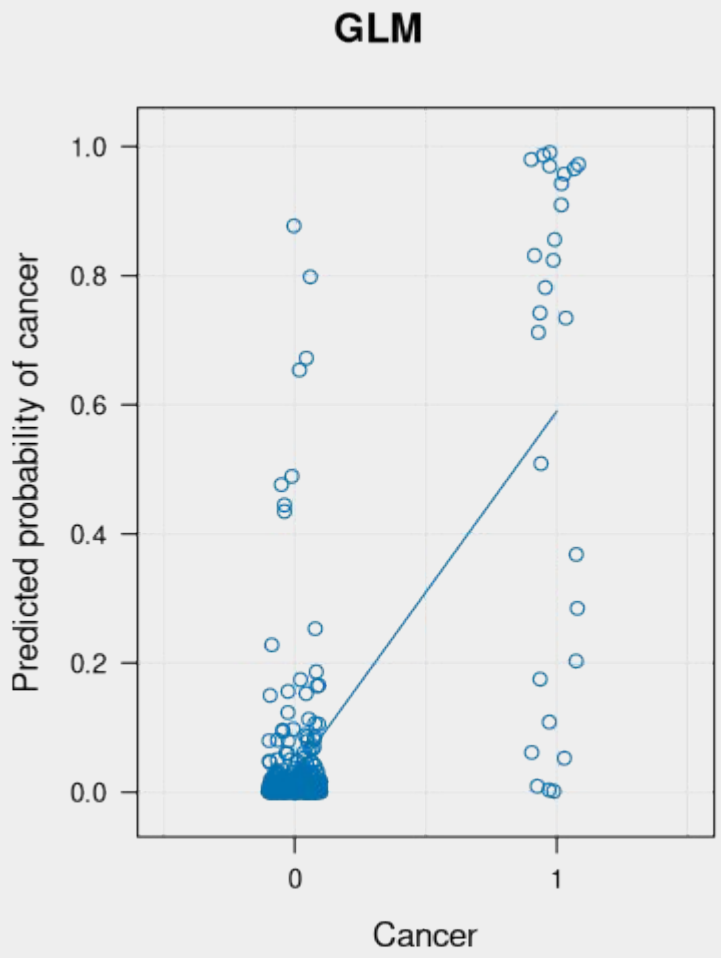
Linear predictors



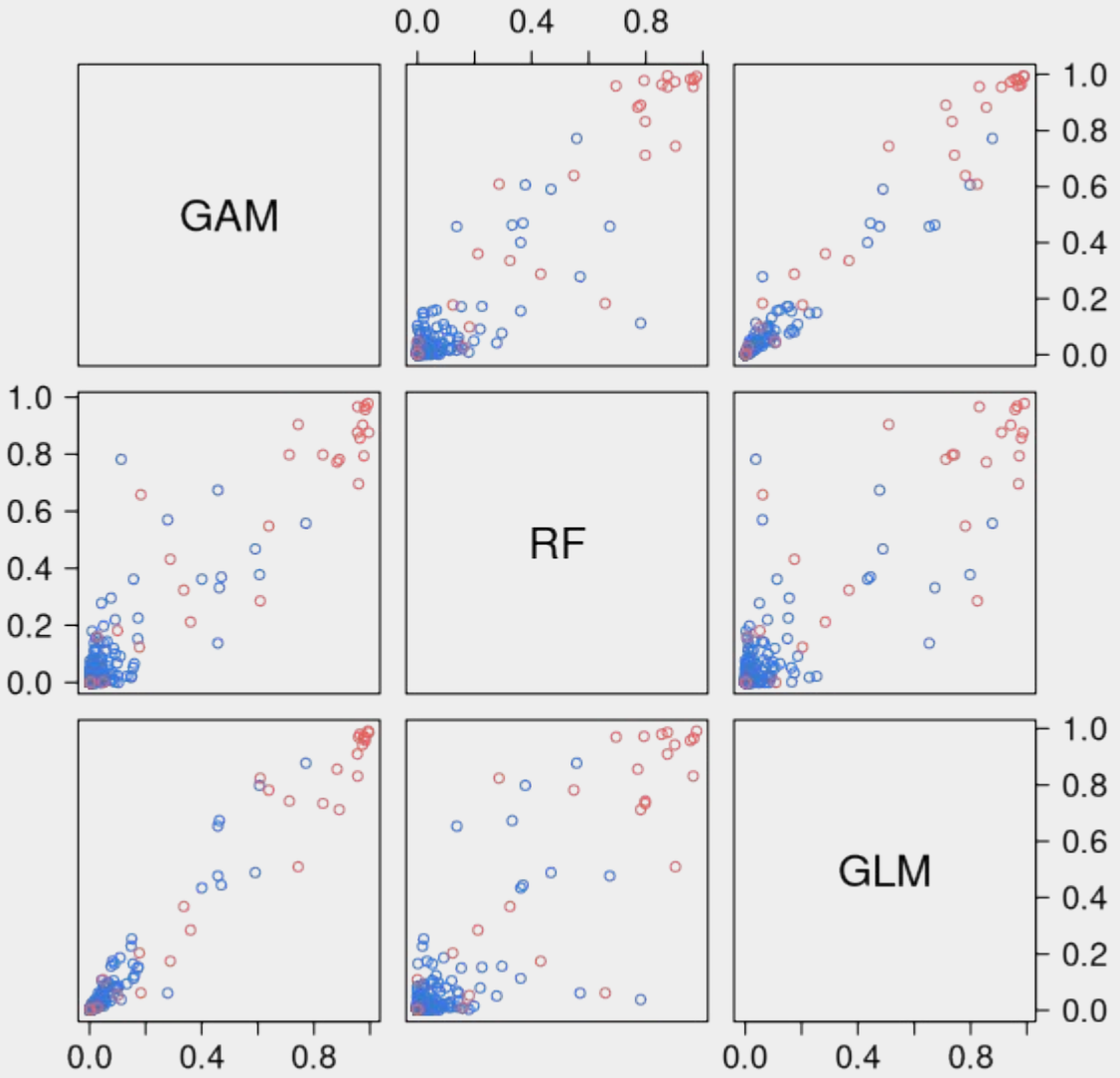
Predictions (LOO)







Predictions (LOO)



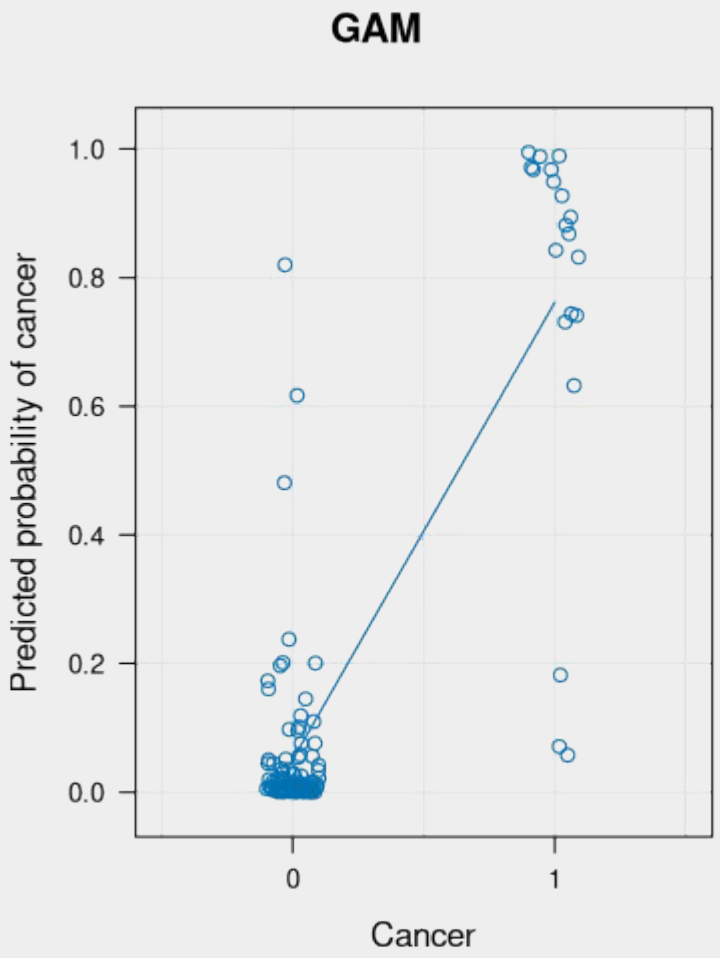
Metrics (LOO)

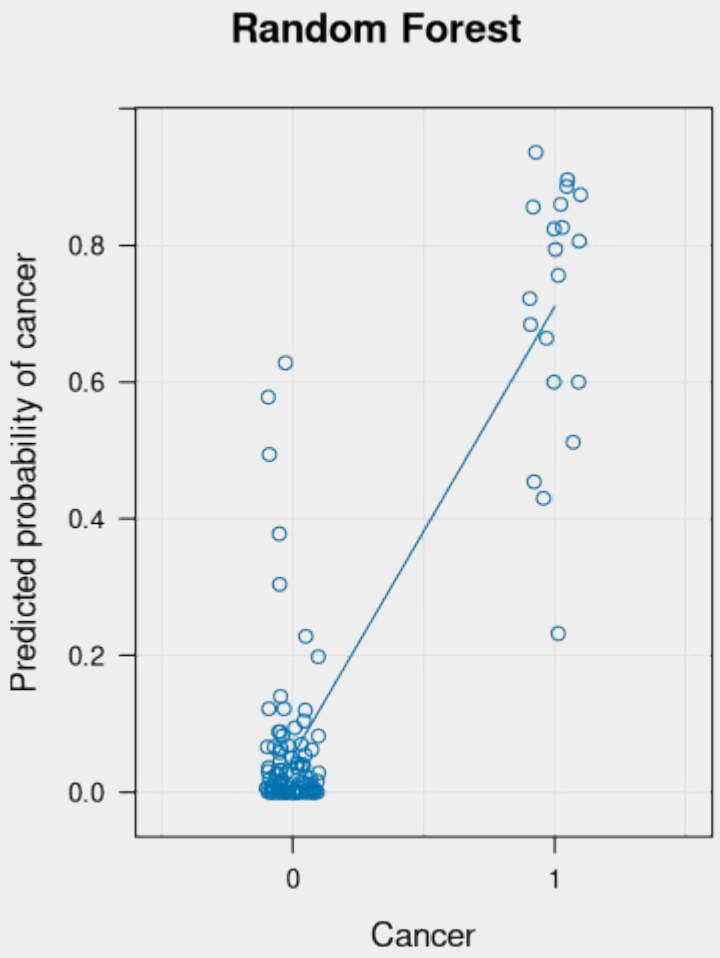
Metrics	GAM	RF	GLM
AUC	0.94	0.92	0.93
Balanced accuracy	0.81	0.81	0.81
Sensitivity	0.63	0.63	0.63
Specificity	0.99	0.99	0.99
PPV	0.85	0.85	0.81
NPV	0.97	0.97	0.97
P4	0.83	0.83	0.82
Log loss	0.56	0.90	0.64
Calibration intercept	0.00	-0.03	0.00
Calibration slope	0.93	0.45	0.93
Expected calibration index	0.05	0.10	0.06

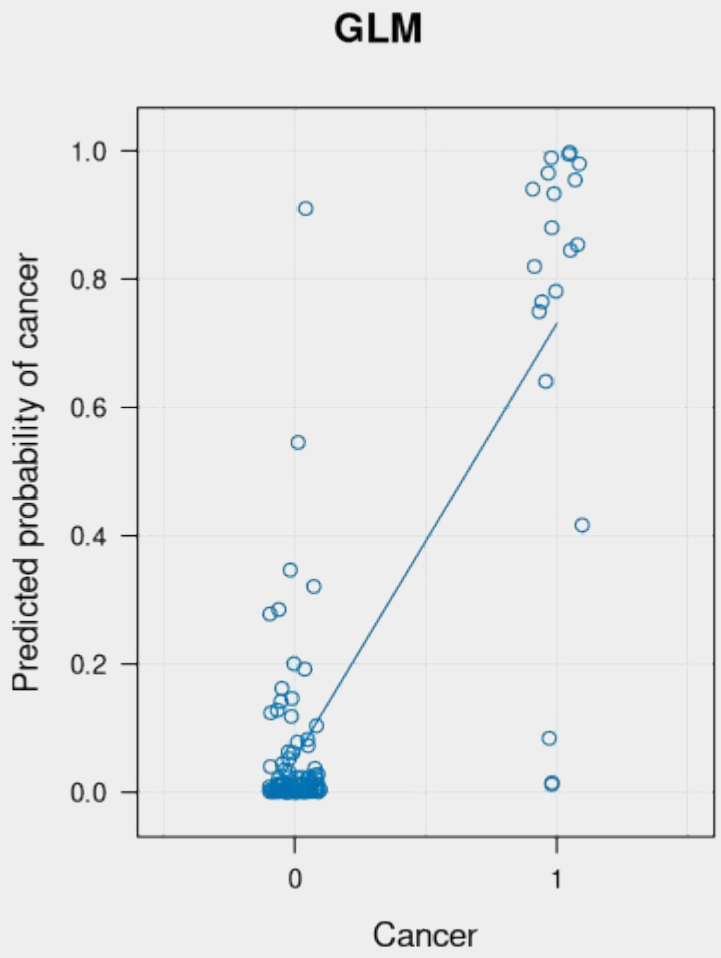
Metrics (LOO)

	Metrics	GAM	RF	GLM
Discrimination	AUC	0.94	0.92	0.93
Classification	Balanced accuracy	0.81	0.81	0.81
	Sensitivity	0.63	0.63	0.63
	Specificity	0.99	0.99	0.99
	PPV	0.85	0.85	0.81
	NPV	0.97	0.97	0.97
	P4	0.83	0.83	0.82
Scoring rule	Log loss	0.56	0.90	0.64
Calibration	Calibration intercept	0.00	-0.03	0.00
	Calibration slope	0.93	0.45	0.93
	Expected calibration index	0.05	0.10	0.06

Predictions (excluded data)



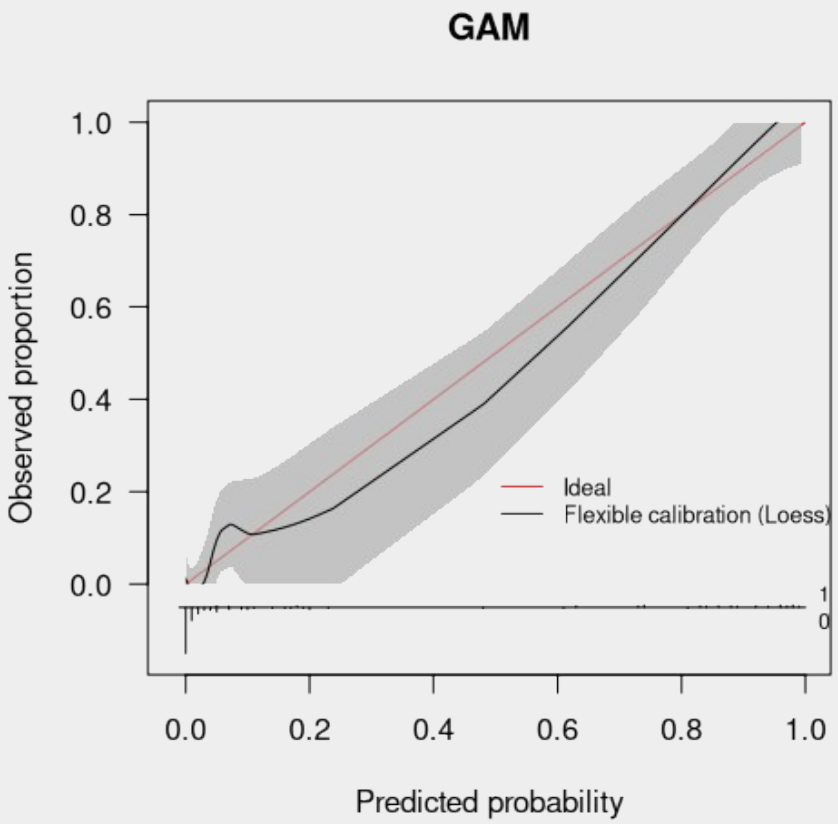


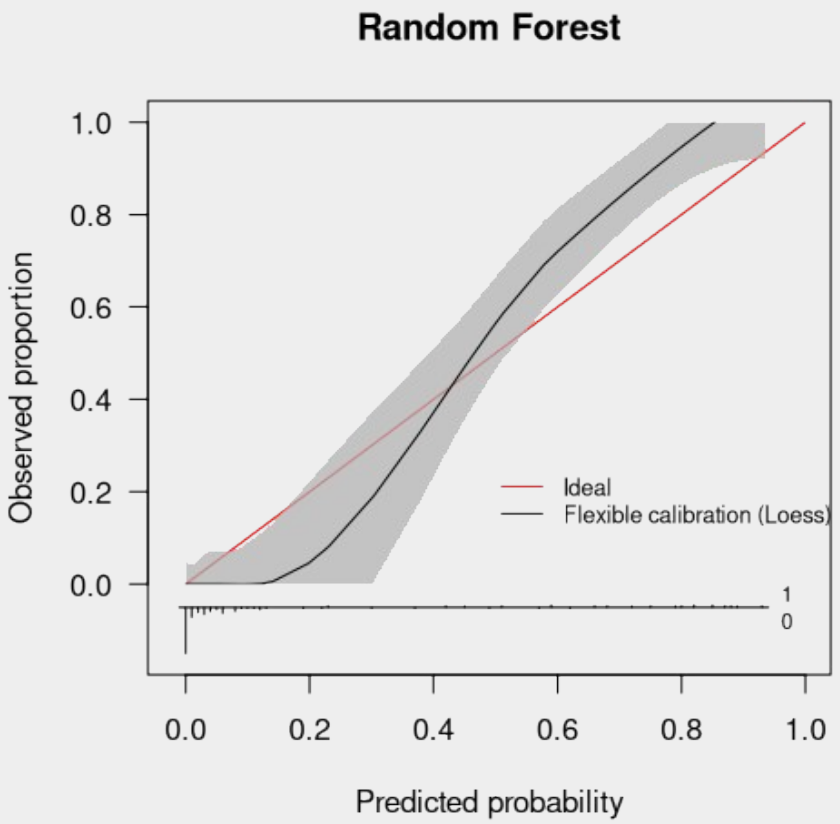


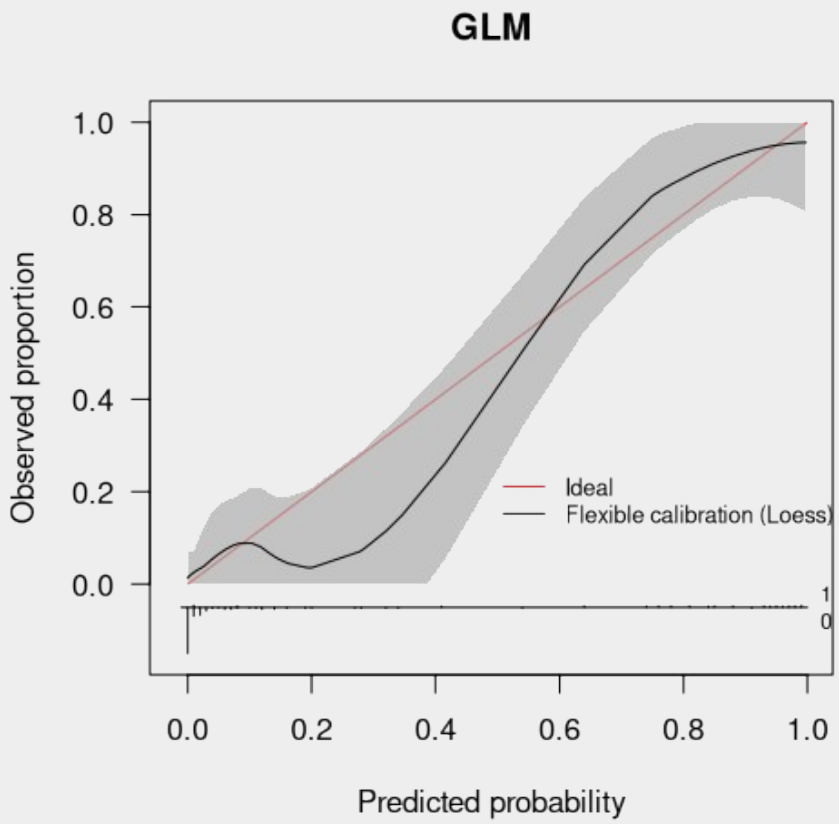
Metrics (excluded data)

	Metrics	GAM	RF	GLM
Discrimination	AUC	0.98	**0.99**	0.95
Classification	Balanced accuracy	**0.93**	0.91	0.89
	Sensitivity	**0.95**	0.85	0.80
	Specificity	0.91	**0.98**	**0.98**
	PPV	0.69	**0.89**	**0.89**
	NPV	**0.99**	0.97	0.96
	P4	0.88	**0.92**	0.90
Scoring rule	Log loss	**0.21**	0.22	0.39
Calibration	Calibration intercept	**-0.04**	0.09	0.06
	Calibration slope	1.12	2.11	**0.89**
	Expected calibration index	**0.10**	0.51	0.29

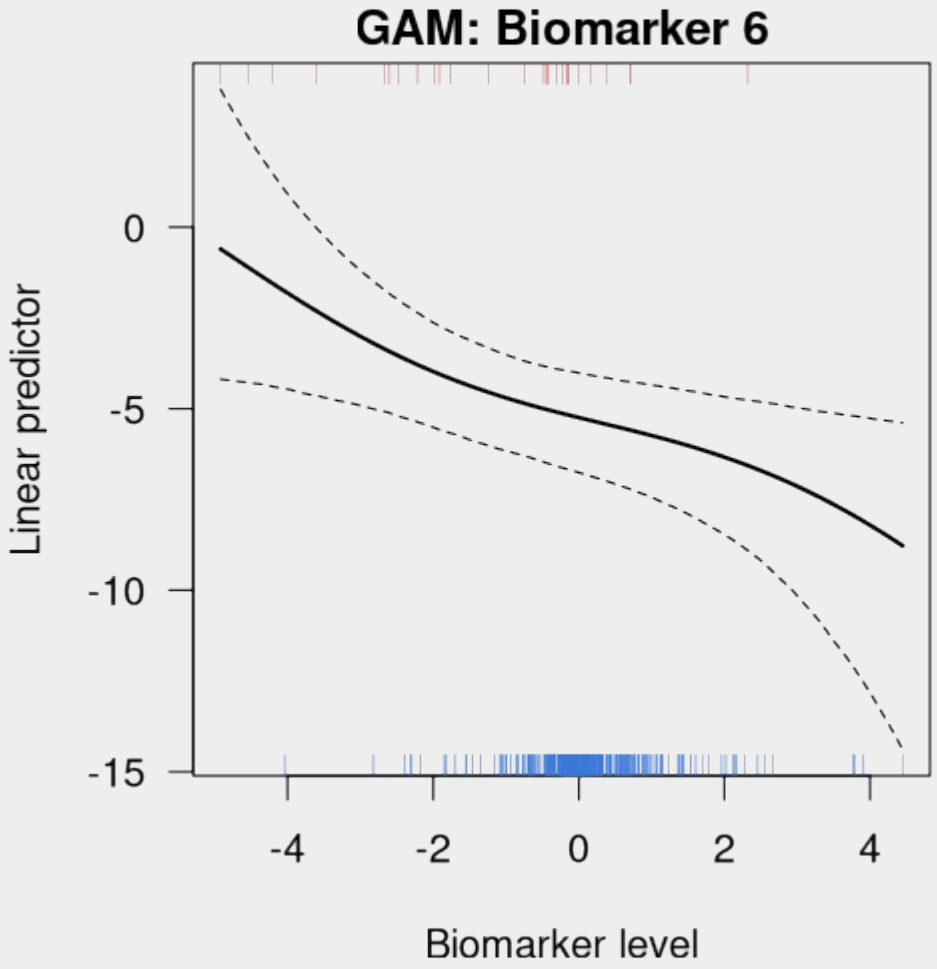
Calibration plots (excluded data)

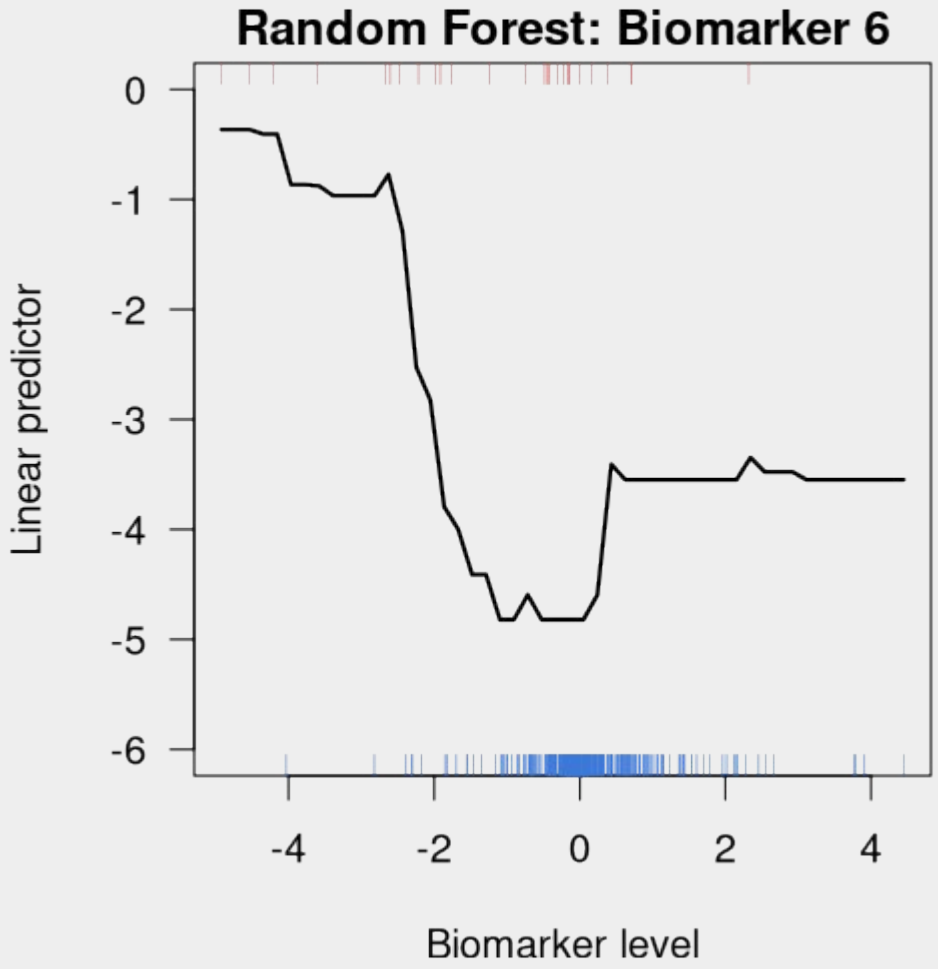






Predictions for Biomarker 6





Drawbacks of GAMs

- Interaction terms harder to incorporate (also reduced interpretability).
- Does not handle cancer sub-types well (e.g. Biomarkers 1 & 2 are high in one sub-type while Biomarkers 3 & 4 are high in another).

Summary

- GAMs provide a class of models that are more flexible than logistic regression.
- They also provide interpretable and biologically plausible predictions (unlike RF, for example).
- They are easy to implement in Stan.